

Original citation:

Zhang, Yan, Holt, Tim A. and Khovanova, N. A.. (2015) A data driven nonlinear stochastic model for blood glucose dynamics. Computer Methods and Programs in Biomedicine.

<http://dx.doi.org/10.1016/j.cmpb.2015.10.021>

Permanent WRAP url:

<http://wrap.warwick.ac.uk/76029>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution 4.0 International license (CC BY 4.0) and may be reused according to the conditions of the license. For more details see: <http://creativecommons.org/licenses/by/4.0/>

A note on versions:

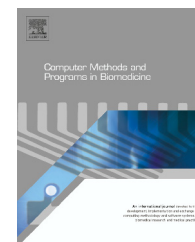
The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: publications@warwick.ac.uk

warwick**publications**wrap

highlight your research

<http://wrap.warwick.ac.uk>

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A data driven nonlinear stochastic model for blood glucose dynamics[☆]

Yan Zhang^a, Tim A. Holt^b, Natalia Khovanova^{a,*}

^a School of Engineering, University of Warwick, UK

^b Department of Primary Care Health Sciences, Oxford University, UK

ARTICLE INFO

Article history:

Received 9 July 2015

Received in revised form

2 October 2015

Accepted 31 October 2015

Keywords:

Nonlinear systems

Data-driven models

Stochastic systems

Blood-glucose dynamics

Diabetes mellitus

System identification

ABSTRACT

The development of adequate mathematical models for blood glucose dynamics may improve early diagnosis and control of diabetes mellitus (DM). We have developed a stochastic nonlinear second order differential equation to describe the response of blood glucose concentration to food intake using continuous glucose monitoring (CGM) data. A variational Bayesian learning scheme was applied to define the number and values of the system's parameters by iterative optimisation of free energy. The model has the minimal order and number of parameters to successfully describe blood glucose dynamics in people with and without DM. The model accounts for the nonlinearity and stochasticity of the underlying glucose–insulin dynamic process. Being data-driven, it takes full advantage of available CGM data and, at the same time, reflects the intrinsic characteristics of the glucose–insulin system without detailed knowledge of the physiological mechanisms. We have shown that the dynamics of some postprandial blood glucose excursions can be described by a reduced (linear) model, previously seen in the literature. A comprehensive analysis demonstrates that deterministic system parameters belong to different ranges for diabetes and controls. Implications for clinical practice are discussed. This is the first study introducing a continuous data-driven nonlinear stochastic model capable of describing both DM and non-DM profiles.

© 2015 The Authors. Published by Elsevier Ireland Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Tight glycaemic control using medication and life style adjustments has proven effective in reducing or delaying complications of DM [1]. There has also been major progress in the development of model predictive control devices, i.e. artificial pancreas systems [2] for type 1 diabetes (T1D), and attempts to automatically control blood glucose variations in people with type 2 diabetes (T2D) [3]. As stated in [2], the vital ingredient

of predictive control algorithms is a model that captures the relationship between glucose excursions, food intake and insulin delivery. The development of such a model represents a major challenge due to limited access to quality data, high cost of equipment for data collection, and, most importantly, the complexity of the underlying systems dynamics. The latter includes four major properties: the glucose–insulin system is open, event-driven, nonlinear and stochastic. These essential properties are rarely incorporated altogether into models of blood-glucose dynamics.

[☆] This work has been supported by EPSRC UK (EP/K02504X/1).

* Corresponding author.

E-mail address: n.khovanova@warwick.ac.uk (N. Khovanova).

<http://dx.doi.org/10.1016/j.cmpb.2015.10.021>

0169-2607/© 2015 The Authors. Published by Elsevier Ireland Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

One of the reasons is that the inherent system complexity often leads to over-complicated models with large numbers of parameters, unsuitable for practical use. Let us briefly summarise two main types of model available in the literature and highlight the advantages of the approach taken in this work.

Phenomenological models [4–8] describe the underlying physiological process (production, distribution, and degradation of glucose and insulin) by compartments, each of which is associated with several differential equations. Although the inherent nonlinearity of the glucose–insulin interaction has been recognised in these models, their application in predictive control is limited by two factors: they are not person specific and contain a high number of system parameters making such models hard to map to low-dimensional data and to validate [9].

Data-driven models [5,10–13] are able to exploit the information hidden in the data and predict glucose concentration without detailed knowledge of the underlying physiological processes [14]. An accepted challenge of such data-driven models is the interpretation of their parameters. They also categorise every poor fit as measurement noise, without taking stochasticity in the data and uncertainty in the model into consideration. In nonlinear systems, noise acts as a driving force; it can radically modify deterministic dynamics, and therefore clearly requires inclusion in the model. This model uncertainty problem can be addressed using probabilistic identification techniques, which operate with parameter distributions rather than single values, and such an approach is taken in this work.

The aim of this work was to find a parsimonious model formulation, i.e. a model of minimal order and with a minimal number of parameters, for postprandial (after a single food intake) glucose dynamics. The model should account for the essential systems properties listed above, and be suitable to support the control strategies in clinical settings. The framework for our approach lies in the area of *data-driven* nonlinear stochastic differential equations modelling. Data-driven models based on differential rather than difference equations have been explored, e.g. in a single case study [15], and for linear systems [16]. It has been shown [15] that postprandial blood glucose excursions can be modelled by a second order linear differential equation in which food intake is treated as a bolus injection of glucose, i.e. an impulsive force. It has also been demonstrated [17] that such linear systems can successfully describe some postprandial glucose excursions in subjects without DM, whereas a strong nonlinear characteristic of the responses was evident for many DM profiles. This highlights the importance of including nonlinear terms in modelling equations as opposed to linear systems approaches.

The structure of the article is as follows. [Section 2](#) gives details on the data available for the analysis. [Section 3](#) explains the methodology for model formulation. [Section 4](#) presents the model, a detailed analysis of systems parameters for profiles with and without DM, and some clinically related interpretation. [Section 5](#) summarises the results, justifies the need for further work and outlines the relevance of the model for DM management.

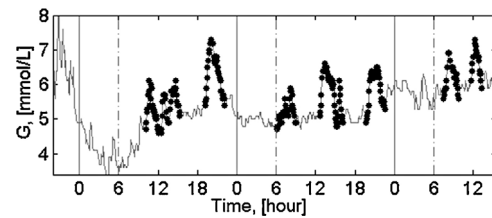


Fig. 1 – Example subcutaneous glucose time series $G(t)$ of a participant from the control group. The solid grey curves represent the measured glucose values and the dots are the values used for modelling of single postprandial peaks. The solid and dashed vertical lines correspond to midnight (0 h) and 6 am, respectively. The first several hours of data (to the left from the first solid vertical line) were excluded from modelling due to the adjustment period of the CGM system.

2. Data description

Datasets from 15 volunteers including five people without DM (control group), four patients with T1D and six patients with T2D was available for the study [17]. Recruitment was purposive to ensure a diverse sample of ages and treatment regimens. Baseline biographical data ([Table A.1](#) in Appendix) were obtained on age, sex, body mass index, type of diabetes, treatment regimen and on recent HbA1c value, indicating the average blood glucose level in the past 2–3 months. Subcutaneous glucose values ([Fig. 1](#)) were taken every 5 min over 72 h using the Medtronic Minimed CGM system [18]. No restrictions were placed on usual daily activities of the participants.

The dotted peaks in the time series ([Fig. 1](#)) represent the postprandial glucose concentration. To avoid mistaking measurement error for genuine postprandial peaks, only distinguishable peaks with height more than 1.1 mmol/L during daytime from 6 am to midnight were selected. The highest peak value for the participant in the control group is just below 8 mmol/L, whereas the highest values for the T1D and T2D patients are greater than 15 mmol/L (not shown).

3. Model and methods

3.1. Model formulation

There are three important characteristics of the blood glucose response to food intake to be taken into account. Firstly, the glucose response is nonlinear as explained in [Section 1](#). Secondly, stochasticity enters into the model in two forms: as measurement error (noise) arising from the device, and/or as dynamical intrinsic noise resulting from factors other than food intake, including physical activity and emotional stress. For patients using insulin, inaccurate estimation of the necessary dose is another factor influencing postprandial excursions. Thirdly, the endocrine system tends to maintain homeostasis, and any deviations of blood glucose from the basal level decay rapidly and return to the pre-disturbed state. The basal glucose level usually demonstrates slow nonstationary dynamics [17]. On the time scales corresponding to a single peak, the variations are small compared with the changes of

baseline values from peak to peak, and therefore the glucose basal level within one peak can be considered as constant. Correspondingly, in our approach, basal level G_b is different for each peak and is set individually according to the experimental dataset. The glucose level over the baseline represents the system's transient response to food intake.

Glucose response to a single food intake in people with and without DM can be considered as the following state-space dynamical model:

$$\frac{d^2}{dt^2}x_t + f_1(x_t)\frac{d}{dt}x_t + f_2(x_t) = F \int \delta(t)dt + \eta_t \quad (1)$$

$$y_t = x_t + \varepsilon_t \quad (2)$$

where

$$f_1(x_t) = \sum_{i=0}^n \theta_{k_i} x_t^i, \quad f_2(x_t) = \sum_{j=1}^m \theta_j x_t^j \quad (3)$$

Eq. (1) is the evolution (system's) equation in which $(x_t, \frac{dx_t}{dt})$ is the state vector of the system [19], and η_t is the dynamic noise. Output x_t of the model (1) has the same meaning and units as $G(t)$; the two different notations are used to distinguish between estimated and measured glucose time series. In the measurement equation Eq. (2), y_t is the output of the model in the presence of measurement noise ε_t . Two stochastic terms, the system noise η_t and the measurement noise ε_t , were modelled as Gaussian-distributed white noise [20] with zero means and intensities (variances) of I_{sys} and I_{meas} , respectively. The deterministic solution of the system (1)–(3), i.e. solution in the absence of both noises, is denoted as x_{td} . In Eq. (1), $F \int \delta(t)dt$ represents an external force (food intake), where F is regarded as an unknown food impact factor, and $\delta(t)$ is the Dirac delta function. The food intake therefore is considered as a bolus injection of glucose at zero time moment [15].

Functions $f_1(x_t)$ and $f_2(x_t)$ play an important role in the suggested model and introduce nonlinearity into the system. As there was no preliminary information on the form of nonlinearity, the polynomial form of the functions was chosen due to its generic nature and the ability to describe a wide variety of solutions x_t . A linear deterministic equation is a particular case of this model when $n=0$, $m=1$, $f_1(x_t) = \theta_{k_0}$, $f_2(x_t) = \theta_1 x_t$ and the stochastic terms are zero: $\eta_t = 0$, $\varepsilon_t = 0$. The linear noise-free system has been considered in [15] and was based on one T2D patient profile.

3.2. Optimal model selection, inference of model parameters and the model's identifiability

The aim was to select the best model structure with the minimal number of parameters for each postprandial peak. To introduce sufficient nonlinearity into the system without overcomplicating the model, three polynomial terms were initially used in both functions f_1 and f_2 . Note that using four or more terms increases the probability of overfitting, rendering the system unstable without improving parameter estimation. While considering the three polynomial terms, some models were disregarded; for example, it is known from the theory

Table 1 – Four model candidates for fitting.

Models	f_1	f_2
M1	$\theta_{k_0} + \theta_{k_1}x_t + \theta_{k_2}x_t^2$	$\theta_1x_t + \theta_2x_t^2 + \theta_3x_t^3$
M2	$\theta_{k_0} + \theta_{k_1}x_t + \theta_{k_2}x_t^2$	θ_1x_t
M3	θ_{k_0}	$\theta_1x_t + \theta_2x_t^2 + \theta_3x_t^3$
ML	θ_{k_0}	θ_1x_t

of dynamical systems that Eqs. (1)–(3) produce unstable solutions for a range of initial conditions if the quadratic term is kept but the cubic term is removed from function f_2 . Using this reasoning, we arrived at the four models presented in Table 1. The first model candidate M1 has full third order polynomial functions f_1 and f_2 . The second candidate M2 has three terms in f_1 and one linear term in f_2 . The third model M3 has three terms in f_2 and one term in f_1 . The three models have been compared with each other and with the linear model (ML).

In order to decide which of the four models best describe each postprandial event and, at the same time, identify corresponding values of deterministic (θ_{k_i} , θ_j) and stochastic (I_{sys} , I_{meas}) parameters, a variational Bayesian learning numerical algorithm was employed [21]. Specifically, the variational Bayesian toolbox [21,22], initially developed for neuroimaging data, was used with the corresponding modifications to make it applicable to our CGM time series. The algorithm accounts for the stochastic nature of the underlying glucose dynamics and also enables us to distinguish the two types of stochasticity: measurement and dynamic noise. This algorithm is flexible and delivers the model parameters in the form of probabilistic distributions rather than fixed values. The final result thereby includes information on uncertainties in parameter estimates. As shown in [21], this approach outperforms standard extended Kalman filtering, the Bethe and Kikuchi family of variational methods and is more robust to model misspecification.

Mathematically, the value of probability $p(y_t|M)$, also referred to as the model evidence, is found by numerically integrating the product of the prior distributions (priors) of the parameters $p(\vartheta|M)$ and the likelihood $p(y_t|\vartheta, M)$ of observing time series y_t given the parameters ϑ and the model M over the distribution of all the parameters ϑ :

$$p(y_t|M) = \int p(\vartheta|M)p(y_t|\vartheta, M)d\vartheta \quad (4)$$

In the Bayesian approach, the choice of priors has a particular importance since it influences the inferred (posterior) parameter distributions. For deterministic parameters, normal prior distributions were considered, which is a commonly used form for an unknown variable. The mean values of the priors for parameters θ_{k_0} and θ_1 of linear terms were chosen in accordance with [15], i.e. set to zero. Also, for consistency, in models M1–M3, mean values of priors for remaining parameters were set to zero. Since there was no prior knowledge about the possible range of parameter values, large variances (10^4) were set, i.e. priors with flat distributions were considered to allow the algorithm searching in a relatively wide region. The large variances provide a weak informative prior, which however converges with iterations towards a narrow final (posterior) distribution. The prior of the food impact factor F was

set to be normally distributed with a mean value of $y_{t_2} - y_{t_1}$, where y_{t_1} and y_{t_2} are the first and second time points of each peak, and variance of 10^4 . The noise precisions, which are inversely proportional to noise intensities and have to take positive values, were modelled by Gamma distributions with both shape and rate parameters set to 0.001 [23].

In Eq. (4), letter ϑ represents the space of all parameters in Eqs. (1)–(3). The integral in (4) is analytically intractable, therefore the Laplace and the mean-field approximations were used to calculate parameter probability distributions $p(\vartheta|M)$. The distribution of each parameter was approximated by the first two moments (mean and variance) known as Laplace approximation [24]. The combined distribution of all the parameters was approximated by the product of individual parameter distributions known as mean-field approximation [24]. The estimated value of $p(\vartheta|M)$ is noted as $q(\vartheta)$, and makes the integral Eq. (4) solvable. The estimated final solution for model evidence $p(y_t|M)$ is smaller than the true value due to decreased entropy [19]. The logarithm of $p(y_t|M)$ is known as ‘free energy’ $\mathcal{F}(q(\vartheta), y_t)$ [19]. This value is maximised during the process of optimising the model for our data and that, among other criteria, defines the goodness of fit. The full procedure for finding the parameters of the best model is referred to as inference.

For each model candidate M (where M is M1, M2, M3 or ML), the probability $p(y_t|M)$ of observing the time series y_t given the model M , was approached by iterative optimisation of the model parameters and the estimated states until a local maximum value of $p(y_t|M)$ was reached. The model with the highest value of $p(y_t|M)$ was selected as the true model.

Note that in the linear case, the selected form of the model guarantees structural identifiability. Indeed, as shown in [25] if a linear stochastic model is stable, controllable and observable, its parameters are identifiable. For linear model ML, both the controllability and observability matrices have full ranks, and we only consider stable solutions. For the nonlinear case, the identifiability is diagnosed via the analysis of the variance of parameters posterior distribution [26], and this is performed automatically in the framework of the variational Bayesian approach.

3.3. Data fitting

Four models (ML, M1, M2 and M3) were compared to achieve the most satisfactory fitting. In order to select the best model, the following two criteria were considered:

- 1 The absolute value of the difference in free energy \mathcal{F} between any two models: $|\Delta\mathcal{F}|$. According to [27], if $|\Delta\mathcal{F}| > 3$, there is strong evidence that the model with the higher value of the free energy is better. This rule is derived from a well known Bayes’ factor which is a measure for comparing the evidences of two Bayesian models [27].
- 2 The decay ratio. Since the glucose dynamics is tightly regulated by the endocrine system, it is a stable process with limited decaying oscillations except for patients in critical conditions or T1D patients who have to maintain the dynamics with insulin injections. Thus, any glucose excursion after food intake should relax to a steady state demonstrating the dynamics of an over-damped system. In

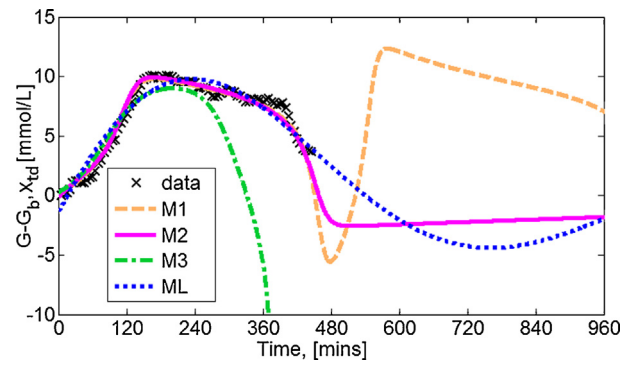


Fig. 2 – Typical outcome for one peak fitting by the four models. G_b is the basal glucose level, G corresponds to the measured glucose values (crosses) and x_{td} corresponds to the deterministic solutions of Eqs. (1)–(2).

the available time series the transient response corresponding to one meal is often superimposed by that of another meal, and consequently small oscillations might be hard to observe. We believe, however, that any model with a decay ratio (ratio of the height of the second peak to the height of the first peak) higher than 0.33 is unrealistic, and should be disregarded.

3.4. Statistical analysis

Statistical analysis of the models’ parameters was performed using the Wilcoxon rank sum test. The null hypothesis of no difference between the groups of interest was tested at the 5% level of significance, and this is presented by p -values.

4. Results and discussion

4.1. Optimal model

Fig. 2 shows (typical) fitting for one postprandial peak by the four suggested models; only deterministic parts x_{td} of the solutions are shown here and in the subsequent Figs. 3 and 5. The most complex nonlinear model M1 produced an unstable result with unrealistic periodic oscillations over time, which was incompatible with the second criterion. Therefore, model M1 was disregarded as unsatisfactory. The results also show that M2, with reduction in the number of polynomial terms of f_2 to one, can successfully constrain the oscillations whereas this is not the case for model M3. M3 was therefore also disregarded, leaving ML and M2 for further consideration.

It is important to note that the linear model ML is a particular case of the nonlinear model M2 for $\theta_{k_1} = 0$ and $\theta_{k_2} = 0$. The use of the two separate models was determined by the feature of the stochastic inference algorithm, which was not able to set the parameters θ_{k_1} and θ_{k_2} at zero value when M2 was applied to a peak with ‘linear’ transient dynamics. In this situation the algorithm converged to a local, rather than the global, maximum of free energy. We were therefore forced to consider two separate models.

Among 130 peaks across all fifteen profiles, 84 peaks were modelled (Table 2) by the linear ML and 46 peaks were fitted

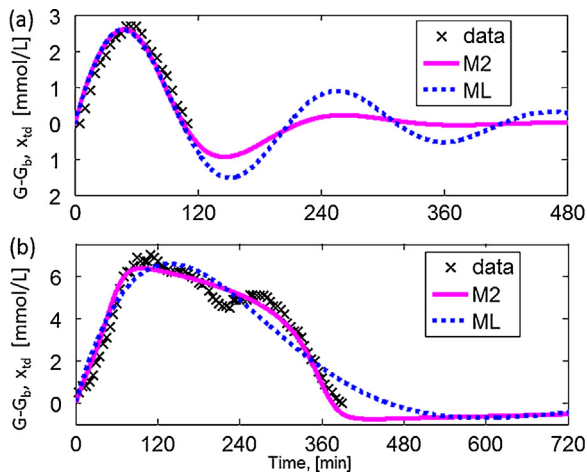


Fig. 3 – The fitting results are shown for: (a) a peak of a T2D profile; (b) a peak of a T1D profile. G_b is the basal glucose level, G corresponds to the measured glucose values (crosses) and x_{td} corresponds to the deterministic solutions of Eqs. (1)–(2).

Table 2 – Summary of peak fitting using ML and M2

Summary	Total peaks	Control	T1D	T2D
Total peaks	130	47	38	45
No. (%) fitted by ML	84	36 (77%)	23 (61%)	25 (56%)
No. (%) fitted by M2	46	11 (23%)	15 (39%)	20 (44%)

by the nonlinear M2. For the control group, ML worked for 77% of the peaks; for the T1D and T2D group, only 61% and 56% of the peaks could be described adequately by ML.

Some examples illustrating how the choice between models ML and M2 was made are presented in Fig. 3. For Fig. 3a, the free energy of ML ($\mathcal{F}_{ML} = 24.0$) is greater than the free energy of M2 ($\mathcal{F}_{M2} = 20.8$) producing $|\Delta\mathcal{F}| = |\mathcal{F}_{ML} - \mathcal{F}_{M2}| = 3.2$. The difference in free energies of the two models is slightly higher than 3, and additionally the linear model does not satisfy the second criterion: the decay ratio is 0.35. This means that the oscillations are considered to be unrealistic, and therefore M2 was chosen as the fitting model. For the peak in Fig. 3b, absolute free energy difference between ML and M2 is 72.6 and free energy of M2 ($\mathcal{F}_{M2} = -1.1$) is much higher than that of ML ($\mathcal{F}_{ML} = -73.7$), indicating a strong nonlinear character in this peak. The linear model is not able to capture the dynamics of this response, which is characterised by the existence of a sub-peak. According to [28], the shape of a peak is dependent on various factors including the macro-nutrients of food intake, gender of the participants and time of food intake. Such double (and multiple) peaks were commonly present in all the profiles; however, subpeaks from T1D and T2D groups had a larger influence over the parameter inference. In DM profiles the peaks are generally higher and occur over a longer time period compared with those from the control group, and as such the effect of the sub-peak dynamics is more distinctive and increases the need for a nonlinear model to capture such features.

Thus, 77% of peaks in the control group can be described by a linear second order model whereas 23% are characterised

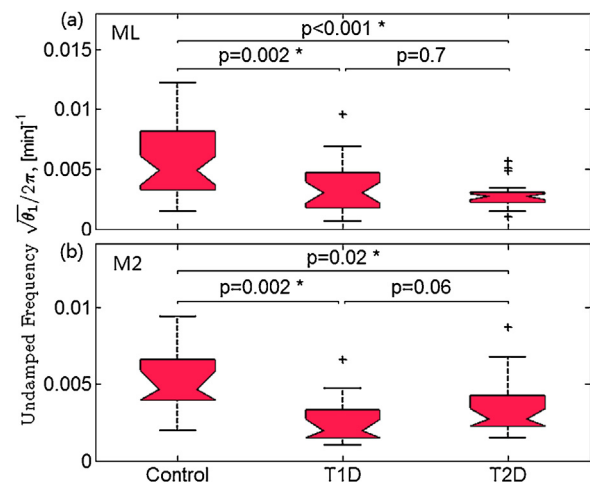


Fig. 4 – Boxplots for parameter $\sqrt{\theta_1}$ obtained from ML (top figure) and M2 (bottom figure).

by nonlinear dynamics. There were more nonlinear peaks in DM groups than in controls: 39% in T1D and 44% in T2D, which emphasises the nonlinear character of response in people with DM. According to [29], the hyper- and hypo-glycaemic regions show greater deviation from linear behaviour, explaining the improved nonlinear fitting for the T2D group.

With respect to stochasticity, the final model for each peak considered above contains both measurement and system noise. However, before the final model was selected as final, two separate models were investigated: one with measurement noise only and another with both measurement and system noise components. The stochastic model with both noise was chosen as the final model as it consistently outperformed the measurement noise only model.

4.2. Parameter comparison between the groups

The deterministic part of ML is represented by two parameters θ_{k_0} [1/min] and θ_1 [1/min²], and of M2 by four parameters θ_{k_0} , θ_{k_1} [L/(mmol min)], θ_{k_2} [L²/(mmol² min)] and θ_1 .

4.2.1. Coefficients of function f_2 and undamped frequency

Functions f_2 have the same structure for both linear and nonlinear models (Table 1), and contain coefficient θ_1 . The square root of this parameter ($\sqrt{\theta_1}$) defines an undamped frequency, at which a system would oscillate in the absence of any driving or damping force. The values of $\sqrt{\theta_1}$ were compared among the three groups (control, T1D and T2D) for models ML and M2 (Fig. 4). If there was a significant difference in medians between any two groups, the p -values are marked by a star. For the peaks that were fitted by ML, the median value of $\sqrt{\theta_1}$ in the control group is significantly higher than the median values in DM groups, whereas there was no statistically significant difference between the two groups of T1D and T2D. Note that the range of parameter values in the control group is also much wider than in the DM groups. For the peaks fitted by M2, the same result is observed: $\sqrt{\theta_1}$ differs significantly between diabetes and controls. Thus, by analysing the undamped

Table 3 – Summary of p -values for coefficients of f_1 in M2.

Parameter	Control-T1D	Control-T2D	T1D-T2D
θ_{k2}	<0.001*	0.02*	0.17
θ_{k1}	<0.001*	0.02*	0.18
θ_{k0}	0.03*	0.02*	0.35

frequency of the models developed it is possible to distinguish between the cases with and without DM.

4.2.2. Coefficients of function f_1 and damping

Functions f_1 characterise damping in the systems and are represented by polynomials of different order for ML and M2. The nonlinear model M2 contains a quadratic nonlinear ‘damping’ function and is characterised by three parameters, whereas the linear model ML contains only one damping coefficient. These inferred coefficients have also been compared across the three groups of controls and diabetes. For the linear damping parameter θ_{k0} of ML, the p -values are as follows: $p=0.03^*$ (control-T1D), $p=0.01^*$ (control-T2D) and $p=0.91$ (T1D-T2D). Statistical analysis for the nonlinear damping parameters of M2 is presented in Table 3. Thus, all three parameters θ_{k2} , θ_{k1} and θ_{k0} of M2 and parameter θ_{k0} of ML differ significantly between diabetes and controls, and, therefore, can be used for differentiating between the groups with and without DM.

It is important to notice that parameter θ_{k0} of the function f_1 defines the system stability of both models, ML and M2. For the system to be stable, the real parts of eigenvalues $\lambda_{1,2}$ of the linear system ML or the linearised system M2 around the steady state $[x_0, \dot{x}_0]^T$ must be negative:

$$\lambda_{1,2} = -\frac{\theta_{k0}}{2} \pm \frac{1}{2} \sqrt{\theta_{k0}^2 - 4\theta_{k1}}. \quad (5)$$

All the inferred parameters θ_{k0} were positive which confirmed the stability of the developed models.

4.2.3. Noise

Parameters of the stochastic terms, i.e the intensities of the system noise I_{sys} [$\text{mmol}^2/(\text{L}^2 \text{min}^4)$] and the measurement noise I_{meas} [mmol^2/L^2] in ML and M2, have also been compared across the three groups.

The intensities of the system noise were of the same order across all three groups for both models without any statistically significant difference: $p>0.05$. The source of the system noise is the aggregate force accounting for other external factors such as physical activity, stress etc. which are not part of the model (we have only considered the impact of food intake on glucose dynamics). The inclusion of the system noise into the analysis provides more flexibility in the model’s structure, but also accounts for its imperfection.

The measurement noise mainly comes from the inaccuracy of the readings from the CGM devices. There were significant differences in I_{meas} values between the three groups for ML. The higher intensities of measurement noise were observed in T1D and T2D groups than in controls ($p<0.001$ and $p=0.003$ correspondingly). This is well justified as the peak glucose levels are generally higher in DM than in the control group, and the CGM devices are less accurate at the larger values. For

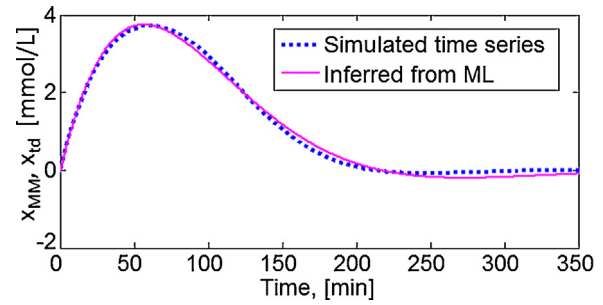


Fig. 5 – (a) Simulated time series from the maximal model, x_{MM} , for a non-DM case without any signs of DM (dotted line), and (b) inferred time series x_{id} using ML (solid line).

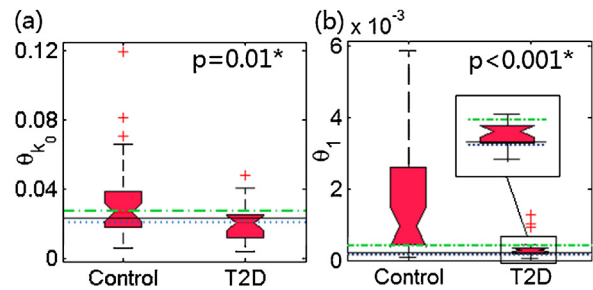


Fig. 6 – Boxplots for (a) θ_{k0} and (b) θ_1 for all measured peaks fitted by ML in our cohort of participants. Horizontal lines mark θ_{k0}^{MM} and θ_1^{MM} for no signs of DM (upper dashed green line), low insulin sensitivity (middle solid line) and impaired β -cell function (lower dotted blue line) cases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

M2, there was no significant difference in measurement noise intensities across the three groups.

4.3. Link between the data-driven model and physiological models. The signs of pre-diabetes

A physiological maximal model (MM) [8] was used to generate time series characterising three non-DM subjects: (i) without any signs of DM, (ii) with low insulin sensitivity, and (iii) with impaired β -cell function. Cases (ii) and (iii) describe potential T2D patients with partially impaired pancreatic function, and are considered as pre-DM cases. The impaired function is compensated by either secreting more insulin or increasing tissue sensitivity, and thus keeping the glucose levels still within the healthy range.

These simulated time-series were used to infer parameters of models (1)–(3) and to select the best model, exactly in the same way as for the measured time-series. It was found that the three cases can be fitted by the linear model ML, and an example result of fitting for case (i) is shown in Fig. 5. Values of the inferred parameter θ_{k0}^{MM} are as follows: (i) 0.027, (ii) 0.023, (iii) 0.021, and the values of θ_1^{MM} are (i) 0.4, (ii) 0.25, (iii) 0.16. They have been compared with θ_{k0} and θ_1 obtained for all measured peaks fitted by ML for our cohort of participants (Fig. 6).

It can be seen that the value of $\theta_{k_0}^{MM}$ for the case without any evidence of pre-DM locates in the interquartile range of the boxplot for the control group (Fig. 6) and in the top quartile for T2D. The values of $\theta_{k_0}^{MM}$ for both pre-DM cases locate in the lower part of the distribution for the control group, but within the top part of the interquartile of the boxplot for T2D. A qualitatively similar result is observed for parameter $\theta_{k_1}^{MM}$ (Fig. 6b): the simulated value for case (i) without DM signs is within the interquartile of the control group distribution and beyond the interquartile of T2D. Similarly, pre-DM simulated values locate within the bottom quartile of the control group and within the interquartile of the T2D range. This clearly shows that the two simulated cases fall into the area between the non-DM and T2D distributions and can be interpreted as pre-DM cases with impaired body functions. Therefore, a close observation of the trend of the parameters θ_{k_0} and θ_{k_1} might facilitate early diagnosis of DM, particularly if such trends identify early abnormalities in dynamics before mean glucose had risen significantly.

Note that together with the important physiological interpretations, this section provides evidence of the robustness of models (1)–(3) and their validation by comparison with an established phenomenological model [8]. The solution of (1)–(3) in Fig. 5b contains only 2 parameters and matches the dynamics of the phenomenological maximal model [8] with 35 parameters (Fig. 5a). Being data-driven, models (1)–(3) take full advantage of CGM data, and, at the same time, reflect the intrinsic characteristics of the glucose–insulin system without detailed knowledge of the underlying physiological mechanisms.

5. Conclusions and limitations of the study

A variational Bayesian method was successfully applied to develop dynamical system equations for glucose time series and to infer system parameters. The results demonstrate a universal nonlinear stochastic model capable of capturing the dynamics of postprandial glucose excursions in diabetes and controls. Deterministic parameters were found to belong to different ranges for people with and without DM, demonstrating the potential for useful clinical applications including early diagnosis of DM, studying the impact of various drugs on the stability of glucose responses as well as control of DM. The form of the model is compatible with phenomenological models which may allow useful physiological parameter interpretations.

This study is limited by a relatively small sample of glucose profiles. However, our results confirm previous findings [15] suggesting an ability to distinguish DM and non-DM cases on the basis of system parameters, with promising interpretations for clinical use. The relationships that we have uncovered between model parameters and clinical phenotypes require further investigation to confirm these associations and elucidate the underlying mechanisms.

Due to the ethically sensitive nature of the research, no interviewees consented to their data being retained or shared. Additional details relating to other aspects of the data are available from the University of Warwick institutional repository: <http://wrap.warwick.ac.uk/74860>.

Appendix A.

Table A.1 – Biometric indices, treatment regimens, HbA1c values and corresponding estimated average blood glucose levels of participants.

Profile no.	Age (years)	Sex	BMI (kg/m ²)	Diabetes status	Treatment regimen	HbA1c (mmol/mol)	Glucose level (mmol/L)
1	57	F	20.5	T1D	basal bolus (glargine plus aspart)	63	10.0
2	27	F	19.2	Control	n/a	n/a	n/a
3	59	F	27.3	Control	n/a	n/a	n/a
4	49	F	21.9	Control	n/a	n/a	n/a
5	32	F	29.4	T1D	Insulin pump	55	9.0
6	74	M	20.5	T2D	Metformin, gliclazide and rosiglitazone	61	9.7
7	66	F	25.9	T1D	Insulin pump	38	6.3
8	75	M	23.4	T2D	Metformin	46	7.6
9	68	F	32.7	T1D	Basal bolus (glargine plus aspart)	48	7.8
10	39	F	21.3	Control	n/a	n/a	n/a
11	61	F	32.6	T2D	Metformin	52	8.4
12	56	M	30.0	T2D	Metformin and gliclazide	68	10.8
13	52	F	44.5	T2D	Metformin and glargine	89	13.8
14	22	F	19.6	Control	n/a	n/a	n/a
15	63	F	27.0	T2D	Newly diagnosed, diet only	42	7.0

REFERENCES

- [1] The Diabetes Control and Complications Trial Research Group, The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus, *N. Engl. J. Med.* (1994) 286–287.
- [2] R. Hovorka, Closed-loop insulin delivery: from bench to clinical practice, *Nat. Rev. Endocrinol.* 7 (7) (2011) 385–395.
- [3] Y. Ruan, H. Thabit, M.E. Wilinska, R. Hovorka, Modelling endogenous insulin concentration in type 2 diabetes during closed-loop insulin delivery, *BioMed. Eng. Online* 14 (1) (2015) 19.
- [4] R.N. Bergman, Lilly lecture 1989. Toward physiological understanding of glucose tolerance. Minimal-model approach, *Diabetes* 38 (12) (1989) 1512–1527.
- [5] P. Palumbo, S. Ditlevsen, A. Bertuzzi, A. De Gaetano, Mathematical modeling of the glucose–insulin system: a review, *Math. Biosci.* 244 (2) (2013) 69–81.
- [6] A. De Gaetano, O. Arino, Mathematical modelling of the intravenous glucose tolerance test, *J. Math. Biol.* 40 (2) (2000) 136–168.
- [7] E.M. Watson, M.J. Chappell, F. Ducrozet, S.M. Poucher, J.W.T. Yates, A new general glucose homeostatic model using a proportional-integral-derivative controller, *Comput. Methods Progr. Biomed.* 102 (2) (2011) 119–129.
- [8] C. Dalla Man, R.a. Rizza, C. Cobelli, Mixed meal simulation model of glucose–insulin system, *Conf. Proc.: Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 1 (10) (2006) 307–310.
- [9] V. Marmarelis, G. Mitsis, in: V. Marmarelis, G. Mitsis (Eds.), *Data-driven Modeling for Diabetes*, ser. *Lecture Notes in Bioengineering*, Springer, Berlin; Heidelberg, 2014.
- [10] J. Florian, R.S. Parker, Empirical Modeling for Glucose Control in Critical Care and Diabetes, *Eur. J. Control* 11 (6) (2005) 601–616.
- [11] Z. Trajanoski, W. Regittnig, P. Wach, Simulation studies on neural predictive control of glucose using the subcutaneous route, *Comput. Methods Progr. Biomed.* 56 (2) (1998) 133–139.
- [12] C. Cobelli, C. Dalla Man, G. Sparacino, L. Magni, G. De Nicolao, B.P. Kovatchev, *Diabetes: models, signals, and control*, *IEEE Rev. Biomed. Eng.* 2 (2009) 54–96.
- [13] A. Makroglou, J. Li, Y. Kuang, Mathematical models and software tools for the glucose–insulin regulatory system and diabetes: an overview, *Appl. Numer. Math.* 56 (3–4) (2006) 559–573.
- [14] E.I. Georga, V.C. Protopappas, D.I. Fotiadis, Glucose Prediction in type 1 and type 2 diabetic patients using data driven techniques, *Knowl. Oriented Appl. Datamin.* (2011) 277–296.
- [15] H.-i. Wu, A case study of type 2 diabetes self-management, *Biomed. Eng. Online* 4 (2005) 4.
- [16] H. Kirchsteiger, R. Johansson, E. Renard, L.D. Re, Continuous-time interval model identification of blood glucose dynamics for type 1 diabetes, *Int. J. Control* (2014) 1–13.
- [17] N.A. Khovanova, I.A. Khovanov, L. Sbano, F. Griffiths, T.A. Holt, Characterization of linear predictability and non-stationarity of subcutaneous glucose profiles, *Comput. Methods Progr. Biomed.* 110 (2013) 260–267.
- [18] M. Minimed, R. City, Continuous glucose monitoring, *Arch. Dis. Childh.* 28 (5) (2005) 150.
- [19] K. Friston, K. Stephan, B. Li, J. Daunizeau, Generalised filtering, *Math. Probl. Eng.* 2010 (2010) 1–34.
- [20] M. Cescon, R. Johansson, E. Renard, A. Maran, Identification of individualised empirical models of carbohydrate and insulin effects on T1DM blood glucose dynamics, *Int. J. Control* 87 (7) (2014) 1438–1453.
- [21] M.J. Beal, Variational algorithms for approximate Bayesian inference, 2003 (PhD dissertation).
- [22] Toolbox SPM. The toolbox is an extension of SPM package developed at UCL, UK. [Online]. Available: <http://www.fil.ion.ucl.ac.uk/spm/>.
- [23] D.J. Spiegelhalter, K.R. Abrams, J.P. Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley, Chichester, 2004.
- [24] J. Daunizeau, K.J. Friston, S.J. Kiebel, Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models, *Physica D: Nonlinear Phenom.* 238 (21) (2009) 2089–2118.
- [25] C. Rangel, J. Angus, Z. Ghahramani, D. Wild, Modeling genetic regulatory networks using gene expression profiling and state-space models, in: *Probabilistic Modeling in Bioinformatics and Medical Informatics SE – 9*, Springer, London, 2005, pp. 269–293.
- [26] K.E. Hines, T.R. Middendorf, R.W. Aldrich, Determination of parameter identifiability in nonlinear biophysical models: a Bayesian approach, *J. Gen. Physiol.* 143 (3) (2014) 401–416.
- [27] R.R.E. Kass, A.E.A. Raftery, Bayes factors, *J. Am. Stat. Assoc.* 90 (430) (1995) 773–795.
- [28] F.Q. Nuttall, M.C. Gannon, J.L. Wald, M. Ahmed, M.A. Frank, Q. Nuttall, Marry C. Gannon, Jean L. Wald, Plasma glucose and insulin profiles in normal subjects ingesting diets of varying carbohydrate, fat, and protein content, *J. Am. Coll. Nutr.* 4 (4) (1985) 437–450.
- [29] R.S. Parker, F.J.D. Iii, J.H. Ward, N.A. Peppas, Robust H Glucose control in diabetes using a physiological model, *Bioeng. Food Nat. Prod.* 46 (12) (2000).